

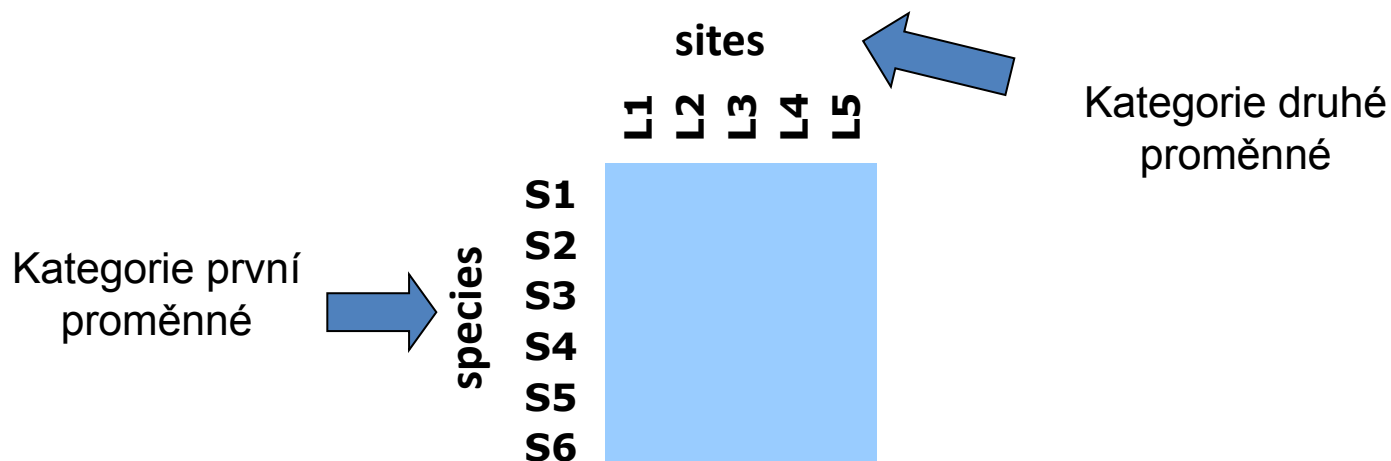
5. cvičení

2.12.2014

Korespondenční analýza

Korespondenční analýza – nástroj pro analýzu vztahů mezi řádky a sloupci kontingenční tabulky => dvě kategoriální proměnné.

Abundance taxonů (nebo počet jakýchkoliv objektů) na lokalitách lze brát jako kontingenční tabulku a mírou vztahu mezi řádky (lokality) a sloupci (taxony) je velikost chi-kvadrátu

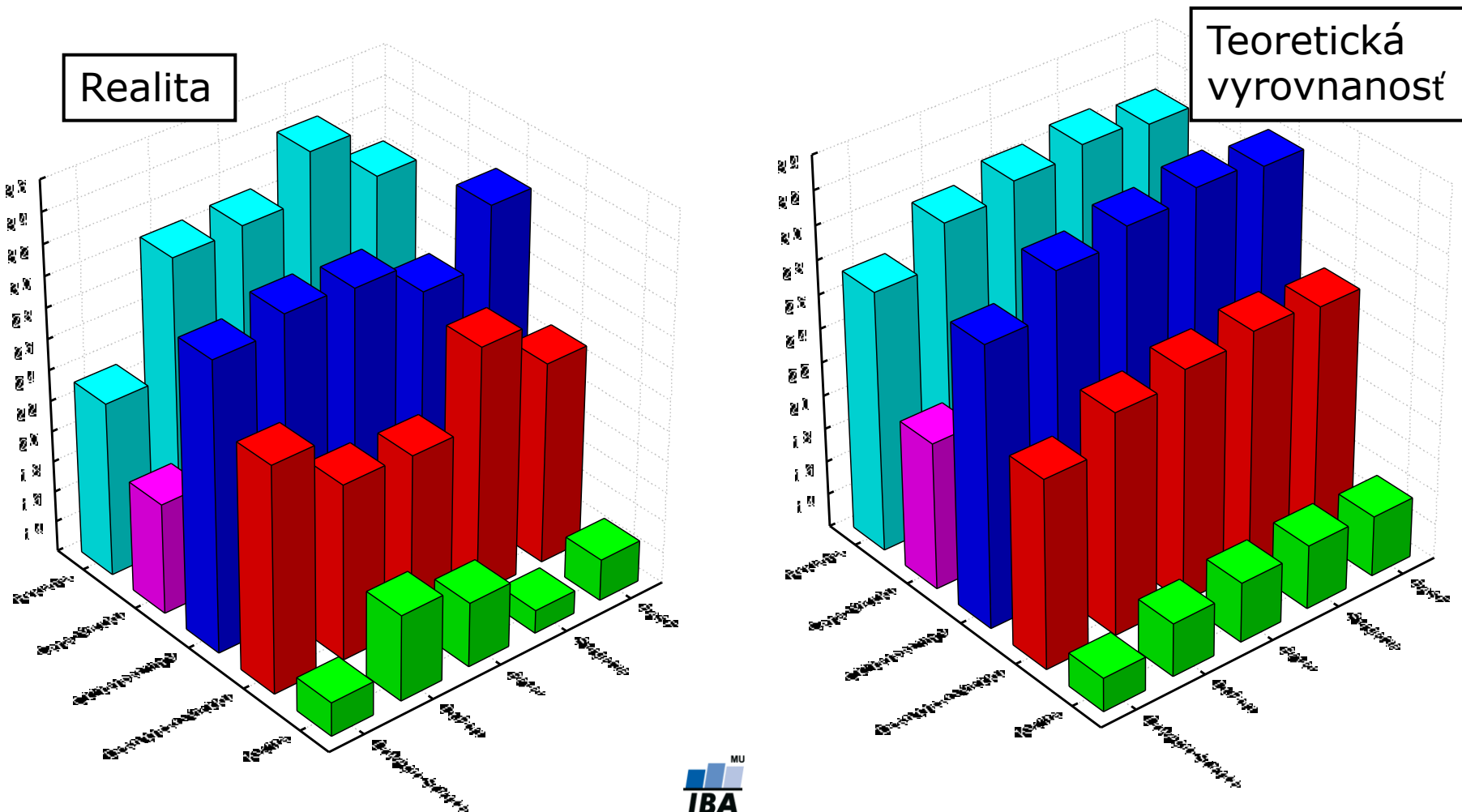


- Základní myšlenkou metody korespondenční analýzy je odvodit indexy (osy), které budou kvantifikovat vztahy mezi řádkovými a sloupcovými kategoriemi. Z těchto indexů můžeme odvodit, která sloupcová kategorie má větší či menší váhu v daném řádku a opačně.
- V grafu interpretujeme relativní pozice bodů řádků a sloupců jako váhy příslouchající danému sloupci a řádku.

Korespondenční analýza

Princip

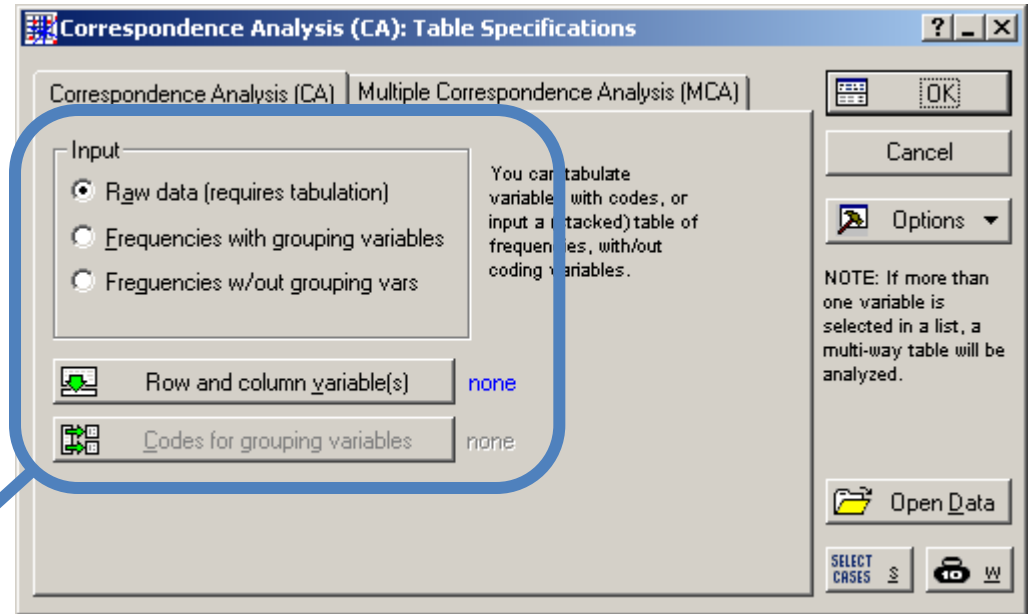
Korespondenční analýza hledá, které kombinace řádků a sloupců hodnocené tabulky nejvíce přispívají k její variabilitě



Korespondenční analýza

Statistics >> Multivariate Exploratory Techniques >> Correspondence Analysis

Podstatou **korespondenční analýzy** je analýza kontingenčních tabulek, tj. tabulek kde průsečíky řádků a sloupců obsahují frekvenci dané kombinace



Nastavení vstupních dat

raw data – proměnné s názvy řádků a sloupců tabulky – frekvence se dopočítají

frekvence s třídící proměnnou – sloupec názvů řádků, názvů sloupců, sloupec s frekvencemi

frekvence bez řídící proměnné – klasická tabulka – řádky X sloupce, na průsečíku frekvence

Korespondenční analýza – výsledky Quick

Koordináty řádků a sloupců v souřadném systému

Výstup všech základních výsledků

Počet rozměrů pro grafy a tabulky

Vybere počet os, vyčerpávajících určitou hodnotu inertia

Correspondence Analysis Results: Dactylogyrus.sta

Number of variables (columns of the table): 10
Number of valid cases (rows of the table): 52

Eigenvalues: ,0922 ,0308 ,0139 ,0120 ,0092 ,0052 ,0028 ,0017 ,0009
Total chi-square=1,92821 df=459 p=1,000

Summary: Row & column coordinates

Print/report summary of results

Number of dimensions:

Number of dimensions: 2

Cumulative contribution to inertia: 90,0 % (or more)

Plots of coordinates:

Row & col., 1D 2D 3D

Cancel

Options

1D, 2D, 3D grafy řádků a sloupců v souřadném systému

Row and column coordinates

Koordináty v ordinačném prostoru CA

Obdoba kvality (\cos^2 a relative inertia pro jednotlivé dimenze

Sloupce nebo řádky (samostatné tabulky)

Column Coordinates and Contributions to Inertia (05_Correspondence.sta)
Input Table (Rows x Columns): 33 x 26
Standardization: Row and Column profiles

Column Name	Column Number	Coord. Dim.1	Coord. Dim.2	Mass	Quality	Relative Inertia	Inertia Dim.1	Cosine2 Dim.1	Inertia Dim.2	Cosine2 Dim.2
druh1	1	0,36515	-0,34050	0,123461	0,091865	0,109609	0,021925	0,049138	0,028166	0,042727
druh2	2	0,36340	-0,26874	0,453295	0,501895	0,059955	0,079724	0,327038	0,064419	0,178857
druh3	3	0,10790	0,21265	0,006517	0,000012	0,012127	0,000101	0,002051	0,000580	0,007961
druh4	4	0,09817	0,07119	0,019551	0,000054	0,012191	0,000251	0,005066	0,000197	0,002689
druh5	5	0,31392	-0,14119	0,010138	0,020060	0,015811	0,0001331	0,020710	0,000403	0,004249
druh6	6	0,14225	0,11137	0,001448	0,000020	0,020962	0,0000039	0,000458	0,000096	0,000761
druh7	7	0,07044	0,00442	0,028602	0,0000129	0,013987	0,0000019	0,003326	0,004879	0,058103
druh8	8	0,32368	-0,2926	0,046343	0,0200116	0,069855	0,006467	0,022783	0,001524	0,003633
druh9	9	0,26360	0,02069	0,008689	0,020002	0,007658	0,000804	0,025843	0,000007	0,000159
druh10	10	0,18414	0,09563	0,105720	0,0900088	0,029033	0,004774	0,000472	0,001903	0,010916

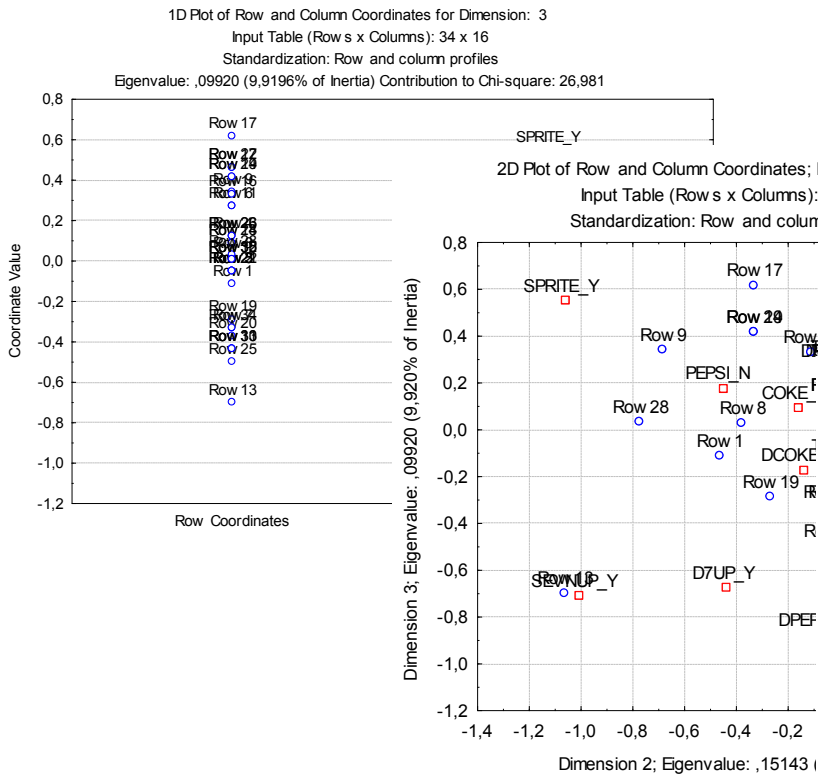
Celkový podíl řádku v tabulce relativních frekvencí (výpočet závisí na nastavení záložky options)

Kvalita zobrazení daného bodu daným počtem dimenzí (proporce bodu k celkové inertii dané počtem dimenzí)

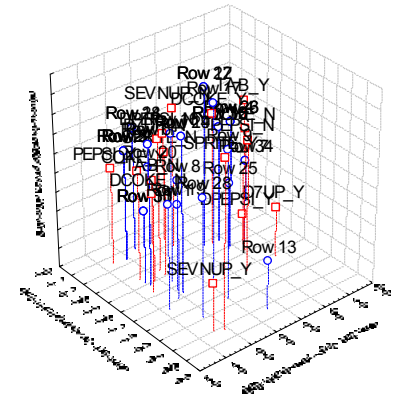
Podíl bodu na celkové inertii (neovlivněno počtem dimenzí)

Grafy CA 1D,2D,3D

Grafy mohou být generovány pro všechny kombinace dimenzí.



3D Plot of Row and Column Coordinates; Dimension: 1 x 2 x 3
 Input Table (Row s x Columns): 34 x 16
 Standardization: Row and column profiles



Grafy obsahují koordináty jak řádků, tak sloupců původní tabulky.

○ Row .Coords
 □ Col.Coords

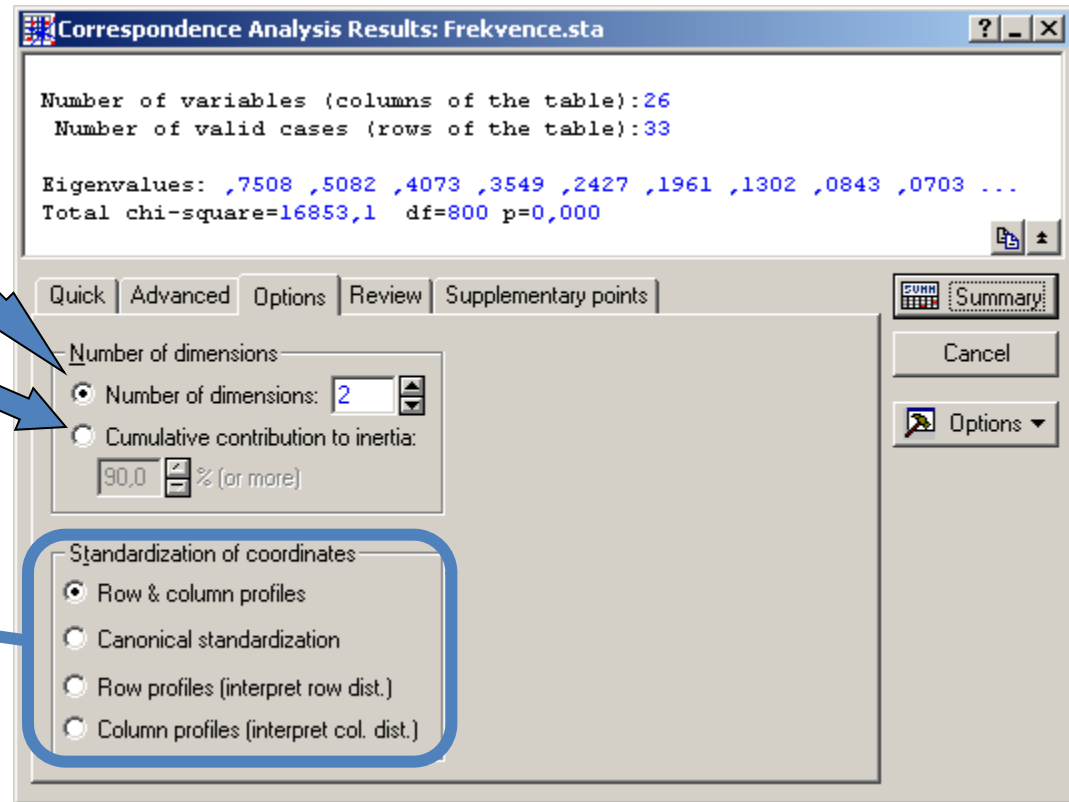
Korespondenční analýza - nastavení

Počet rozměrů pro grafy a tabulky

Vybere počet os, vyčerpávajících určitou hodnotu inertia

Způsob standardizace koordinátů

1. Interpretace vzdáleností v rámci řádků i sloupců
2. Kanonická standardizace
3. Interpretace jen v rámci řádků
4. Interpretace jen rámci sloupců



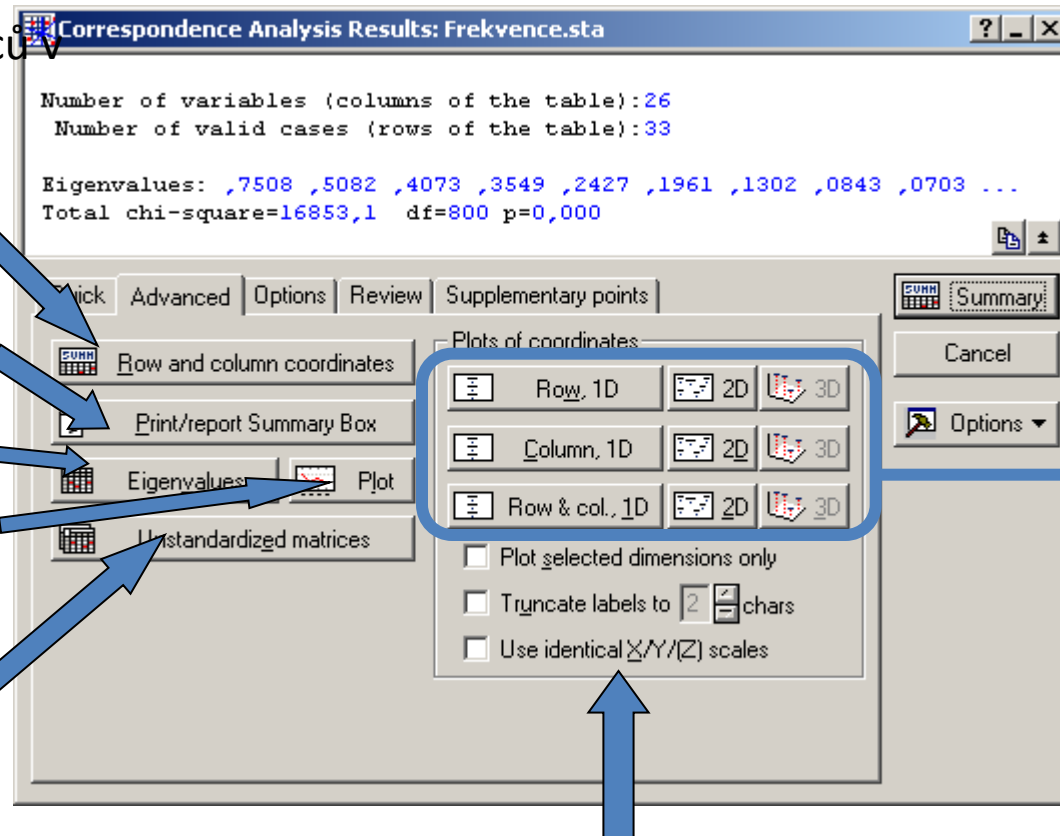
Korespondenční analýza– výsledky Advanced

Koordináty řádků a sloupců v
souřadném systému

Výstup všech
základních výsledků

Eigenvalues ~ inertia
“vysvětlená“ sloupce +
graf

Tabulka frekvencí výskytu
+ nestandardizované
koordináty



1D, 2D, 3D
grafy řádků a
sloupců v
souřadném
systému

Nastavení grafů – výběr os, zkrácení popisek, identické
měřítko

% inertia a kumulativní inertia
vybraná dimenzí

eigenvalue

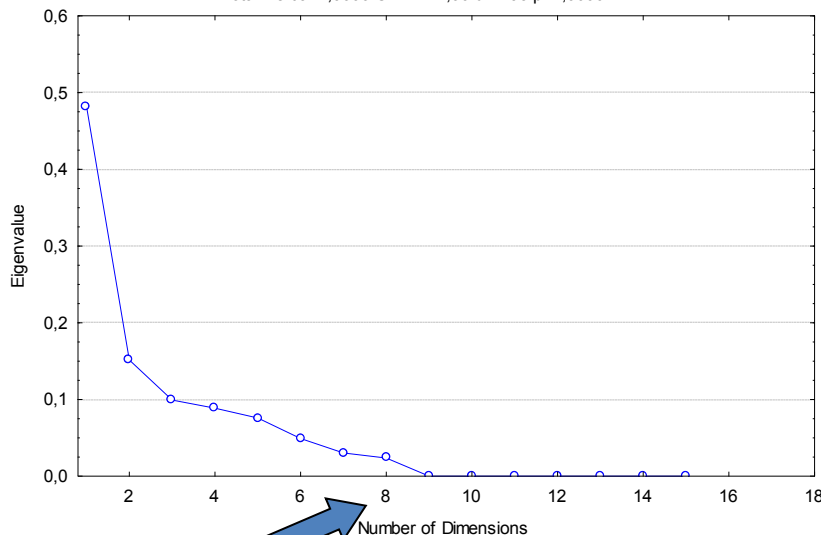
Eigenvalues and Inertia for all Dimensions (Beverage.sta
Input Table (Row s x Columns): 34 x 16
Total Inertia=1,0000 Chi2=277,50 df=495 p=1,0000

Number of Dims.	Singular Values	Eigen-Values	Perc. of Inertia	Cumulativ Percent	Chi Squares
1	0,694078	0,481745	48,17449	48,1745	131,0346
2	0,389141	0,151430	15,14304	63,3175	41,1891
3	0,314954	0,099196	9,91959	73,2371	
4	0,297800	0,088685	8,86851	82,1056	
5	0,275500	0,075900	7,59004	89,6956	
6	0,221580	0,049098	4,90976	94,6054	
7	0,174100	0,030311	3,03109	97,6365	
8	0,153736	0,023635	2,36348	100,0000	
9	0,000000	0,000000	0,00000	100,0000	
10	0,000000	0,000000	0,00000	100,0000	
11	0,000000	0,000000	0,00000	100,0000	
12	0,000000	0,000000	0,00000	100,0000	

Vysvětlený χ^2

Jednotlivé dimenze

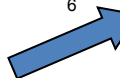
Plot of Eigenvalues
Input Table (Row s x Columns): 34 x 16
Total Inertia=1,0000 Chi2=272,00 df=495 p=1,0000



eigenvalue



Počet dimenzí



Korespondenční analýza analyzuje kontingenční tabulky, k původní tabulce frekvencí je vytvořena tabulka očekávaných frekvencí a tyto dvě tabulky jsou pomocí χ^2 srovnány, analýza hledá takové nové dimenze, které vyčerpávají maximální část celkové χ^2 hodnoty (tzv. inertia)

Matrix of Relative Frequencies (05_Correspondence.sta)
Input Table (Rows x Columns): 33 x 26

	druh1	druh22	druh23	druh24	druh25	druh26	Total
HEL	0,000000	0,000000	0,000000	0,000000	0,000000	0,000000	0,016655
TYM	0,001448	0,000000	0,000000	0,000000	0,000000	0,000000	0,005612
KYJ	0,000000	0,000000	0,000000	0,000000	0,000000	0,000000	0,007698
MOR	0,000000	0,000000	0,000000	0,000000	0,000000	0,000000	0,001448
KYJ	0,001086	0,000000	0,000000	0,001267	0,000000	0,000000	0,178538
MOR	0,000362	0,000000	0,000000	0,000000	0,000181	0,000000	0,034938
KYJ	0,000000	0,000000	0,000000	0,000000	0,000000	0,000000	0,003077
MOR	0,033128	0,000000	0,000000	0,000000	0,000000	0,000000	0,051774
Total	0,123461	0,000905	0,000362	0,002896	0,002715	0,000161	1,000000

řádky

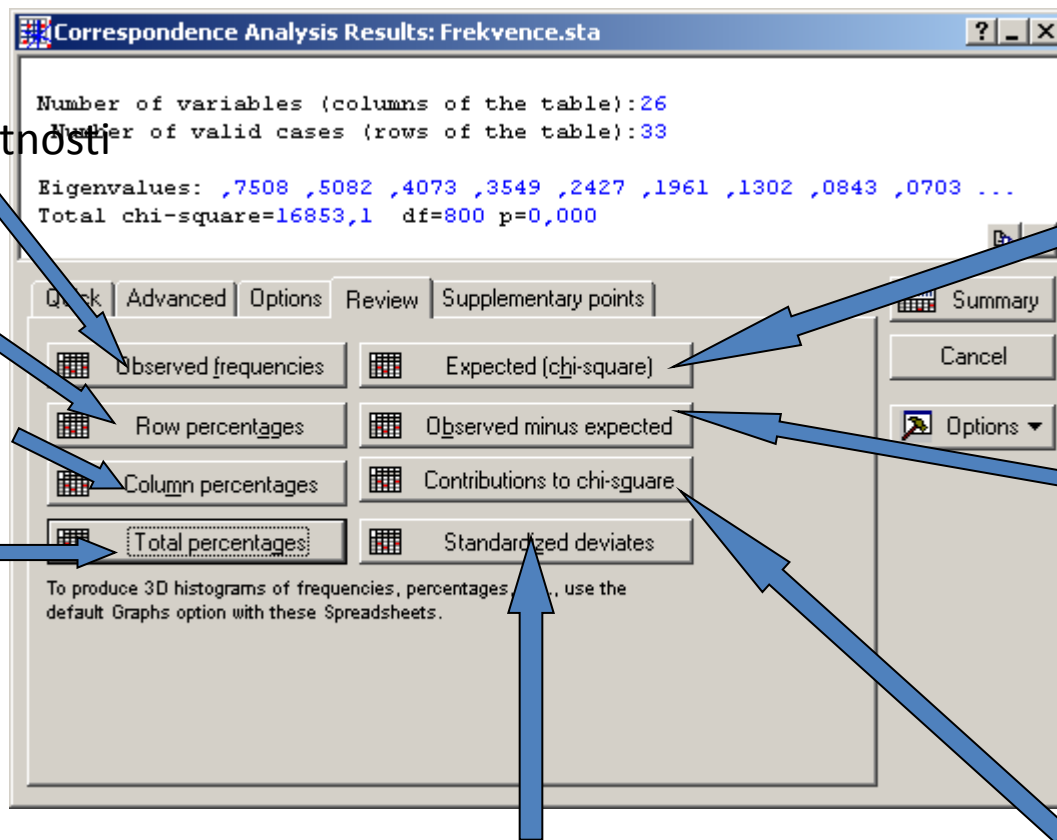
relativní frekvence bodů v původní matici

sloupce

Podíly řádků a sloupců

suma celé matice

Korespondenční analýza– přehledy



Pozorované četnosti

Očekávané četnosti

Rozdíl pozorovaných a očekávaných frekvencí

Podíly v řádcích

Podíly v sloupcích

Podíly v celé tabulce

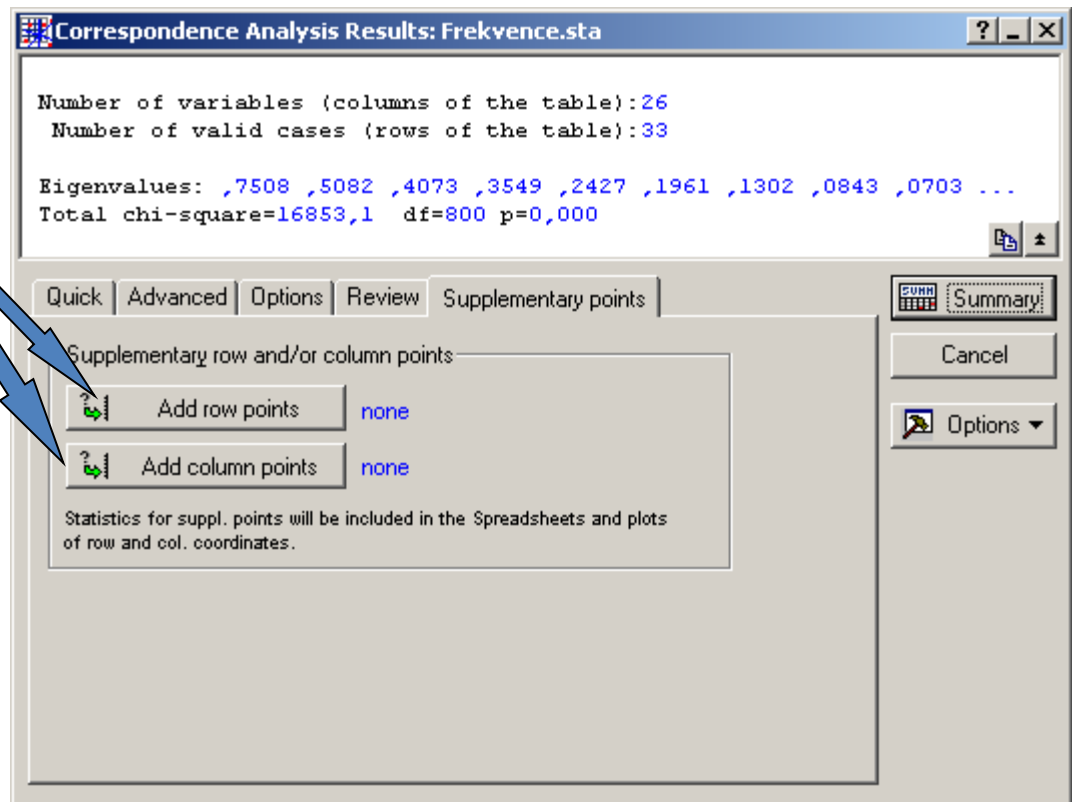
Standardized deviates – odmocnina vlivu na Chi – square + doplnění znaménka

Vliv jednotlivých položek tabulky na celkový Chi-square

K výsledkům analýzy je možné přidat další řádky nebo sloupce, jejichž pozice v souřadném prostoru se spočítají na základě CA, ale její výpočet neovlivní (obdobu supplementary variables a ne-active cases u PCA)

Přidání dalších řádků nebo sloupců

Přidají se do analýzy na základě již spočítaných parametrů



Nemetrické mnohorozměrné škálování (NMDS)

Cílem analýzy je zobrazit pozorované podobnosti nebo nepodobnosti (vzdálenosti) mezi zkoumanými objekty v euklidovském prostoru; zachovává pouze pořadí vzdáleností

Pomocí NMDS můžeme analyzovat nejenom korelační matice (tak jako je tomu v PCA) ale i jakoukoliv jinou matici podobnosti/nepodobnosti

Výhody:

Neparametrická ordinace – robustnější k odlehlým hodnotám

Dá se použít před použitím nehierarchického shlukování k -průměrů (v případech kdy není možné použít euklidovské vzdálenosti)

Nevýhody:

Nutnost specifikovat počet dimenzí předem

Těžko interpretovatelné výsledky v některých případech

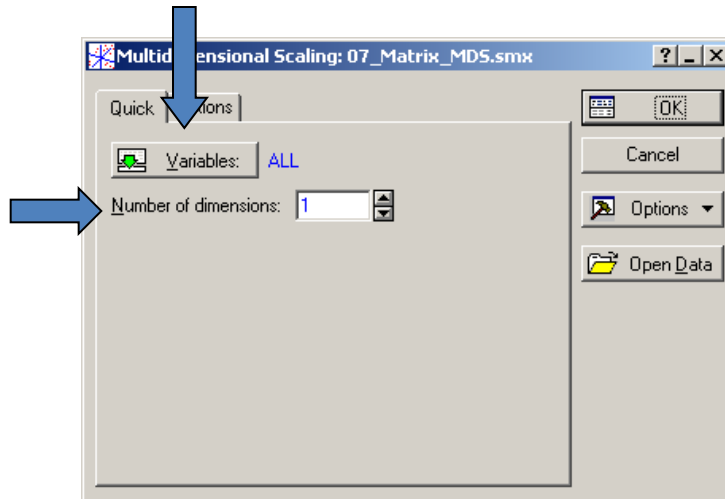
Multidimensional Scaling

Statistics >> Multivariate Exploratory Techniques >> Multidimensional Scaling

Multidimensional scaling dokáže na základě asociační matice s libovolnou metrikou vytvořit její Euklidovskou reprezentaci (příklad: na základě tabulky vzdáleností měst vytvoří mapu).

Výběr parametrů (vstupní soubor musí mít formát asociační matice)

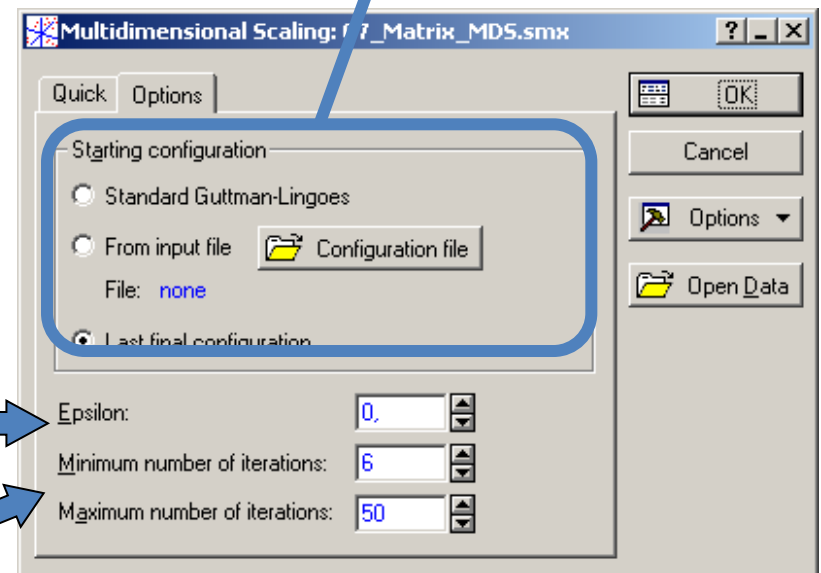
Počet dimenzí k extrakci



Vzdálenosti menší než jsou považovány za 0

Počty iterací

Počáteční konfigurace



Multidimensional Scaling - výpočet

Multidimensional scaling může sloužit pro přípravu podkladů pro k-means clustering pokud nemůžeme na naše data použít Euklidovskou vzdálenost. Metoda je výpočetně velmi náročná.

iter.	[dim=1]	D-star	D-star	D-hat	d-hat		
s: t:	cosin	step	raw stress	alienation	raw stress	stress	
59	1	,758	,081			,0000005	,0000214
60	1	,518	,051			,0000004	,0000197
61	1	,672	,055			,0000004	,0000183
62	1	,891	,099			,0000003	,0000159
63	1	,826	,098			,0000002	,0000141
64	1	,424	,050			,0000002	,0000129
65	1	,515	,043			,0000002	,0000122
66	1	,901	,094			,0000001	,0000107
67	1	,942	,141			,0000001	,0000088
68	1	,604	,069			,0000001	,0000080
69	1	,262	,041			,0000001	,0000075
70	1	,770	,063			,0000001	,0000068
71	1	,939	,122			,0000000	,0000058
72	1	,802	,098			,0000000	,0000051
73	1	,360	,048			,0000000	,0000047
72	*			,0000000	,0000057	,0000000	,0000047
56	1	,624	,054			,0000010	,0000300
57	1	,795	,074			,0000008	,0000271
58	1	,850	,096			,0000006	,0000238

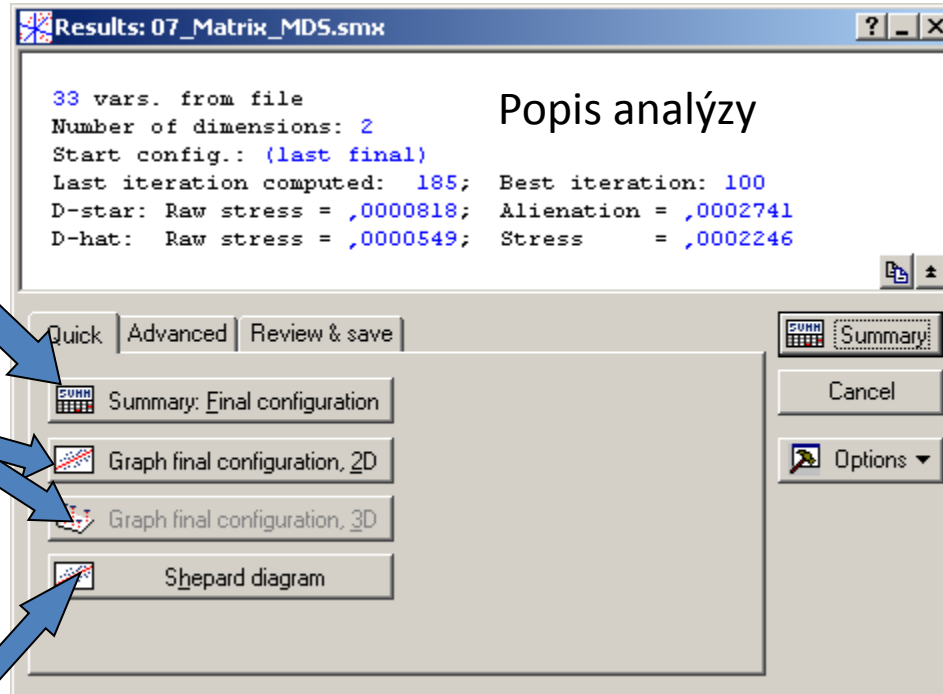
Parametry měnící se při přepočtech

Multidimensional Scaling – výsledky Quick

Výstup nových dimenzí + charakteristiky

Výstupní 2D a D graf

Shepard diagram ~ věrnost reprezentace



Multidimensional Scaling – výsledky tabulky

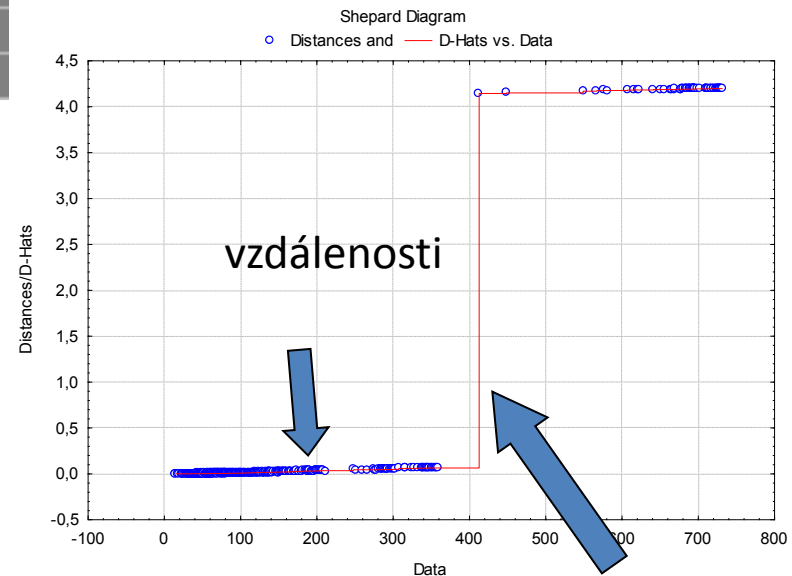
objekty

Final Configuration (07_Matrix_MDS.smx)
 D-star: Raw stress = ,0000818; Alienation = ,0002741
 D-hat: Raw stress = ,0000549; Stress = ,0002246

	DIM. 1	DIM. 2
HEL	-0,254837	0,000981
HVE	-0,254777	0,000982
MEL	-0,254272	0,000491
ROH	-0,255098	-0,002506

Nové dimenze

Shepard diagram

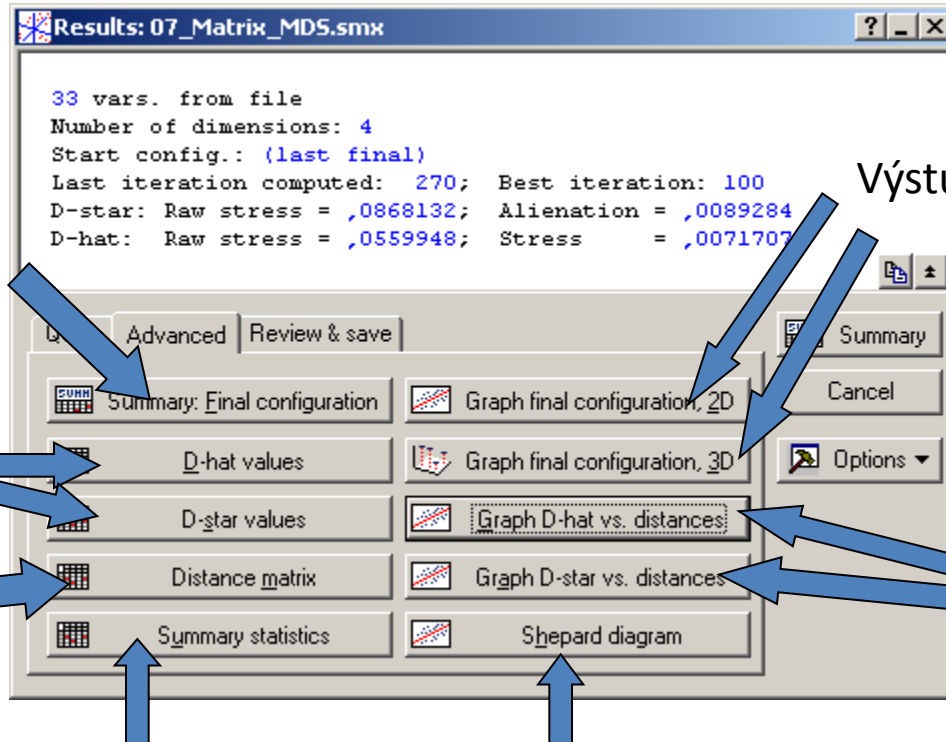


Stress – měřítko reprezentace, čím nižší, tím lepší reprezentace

Alienation – cizost, čím nižší, tím lepší reprezentace

D-hat ~ průběh vzdáleností při dobré reprezentaci

Multidimensional Scaling – výsledky Advanced



Výstup nových dimenzí + charakteristiky

Výstupní 2D a 3D graf

D-hat, D-star

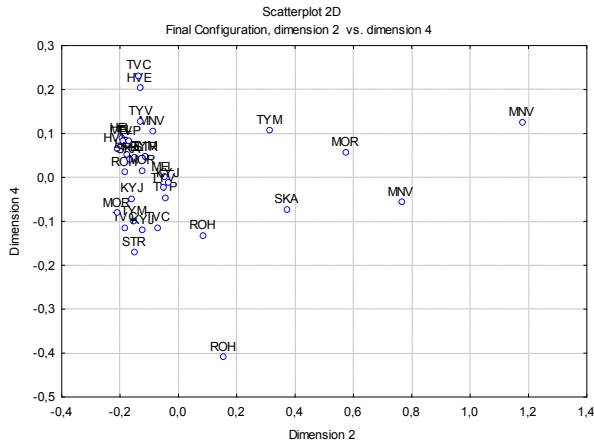
D-hat, D-star versus reprodukováná vzdálenost ~ věrnost reprodukce

Malice vzdáleností (reprodukována)

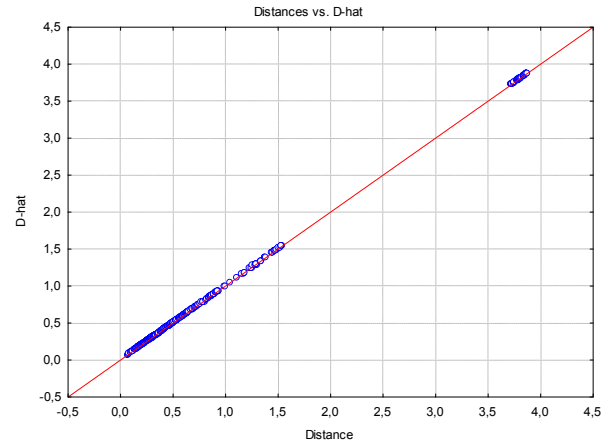
Sumární hodnoty (reprodukováná vzdálenost, D-hat, D-star)

Shepard diagram

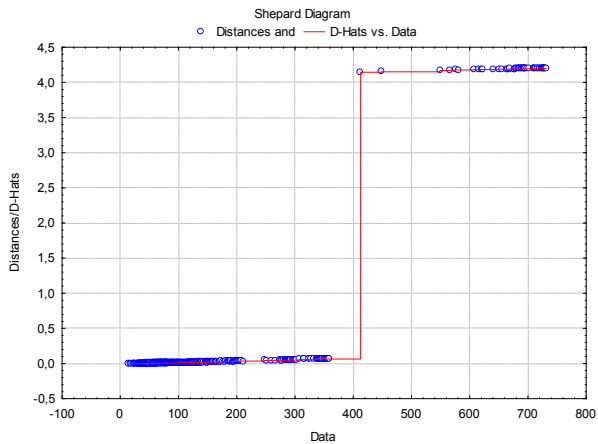
2D graf



Vzdálenosti x D-hat (D-star)



Shepard diagram



3D graf

